KoELECTRA를 활용한 챗봇 데이터의 혐오 표현 탐지

신민기[O]1,2, 진효진2, 송현호1,2, 최정회3, 임현승3, 차미영2,1

1KAIST 전산학부, 2IBS 데이터 사이언스 그룹, 3심심이 주식회사

mingi.shin@kaist.ac.kr, tesschin@ibs.re.kr, hyun78@kaist.ac.kr,

sijay@simsimi.com, hslim@simsimi.com, meeyoungcha@ibs.re.kr

Hate Speech Detection in Chatbot Data Using KoELECTRA

Mingi Shin[O]1,2, Hyojin Chin2, Hyeonho Song1,2,

Jeonghoi Choi3, Hyeonseung Lim3, Meeyoung Cha2,1

1KAIST School of Computing, 2IBS Data Science Group, 3SimSimi Inc.

요 약

챗봇과 같은 대화형 에이전트 사용이 증가하면서 채팅에서의 혐오 표현 사용도 더불어 증가하고 있다. 혐오 표현을 자동으로 탐지하려는 노력은 다양하게 시도되어 왔으나, 챗봇 데이터를 대상으로 한 혐오 표현 탐지 연구는 여전히 부족한 실정이다. 이 연구는 혐오 표현을 포함한 챗봇-사용자 대화 데이터 35만 개에 한국어 말뭉치로 학습된 KoELETRA 기반 혐오 탐지 모델을 적용하여, 챗봇-사람 데이터셋에서의 혐오 표현 탐지의 성능과 한계점을 검토하였다. KoELECTRA 혐오 표현 분류 모델은 챗봇 데이터셋에 대해 가중 평균 F1-score 0.66의 성능을 보였으며, 오탈자에 대한 취약성, 맥락 미반영으로 인한 편향 강화, 가용한 데이터의 정확도 문제가 주요한 한계로 포착되었다. 이 연구에서는 실험 결과에 기반해 성능 향상을 위한 방향성을 제시한다.

주제어: KoELECTRA, Transformer, 혐오 표현 탐지, 욕설, 챗봇, 대화 데이터

1. 서론

최근 챗봇, 스마트스피커와 같은 상업용 대화형 에이전트의 보급률이 폭발적으로 증가함과 동시에 대화형 에이전트를 대상으로 한 비정상 커뮤니케이션도 증가하고 있다. AI를 대상으로 한 비윤리적 발언과 혐오 표현 문제는, AI를 타깃으로 한 폭력이 사람에게로 그 표적이 옮겨가거나[1], 이루다 이슈와 같이[1] AI의 비윤리적 발언으로 인해 제조사가 장기간 준비한 서비스를 3주 만에 중단해야 하는 등 사회적 파급력이 큰 문제이다. 그러나, 혐오 표현이 담긴 가용 학습 데이터 부족, 기술적 어려움 및 윤리적인 기준 미확립 등의 문제로 관련 연구가 미비한 실정이다.

온라인상의 차별과 혐오 발언 증가로 인해 인공지능을 활용하여 데이터 속 혐오 표현을 탐지하는 연구는 NLP와 AI 분야의 중요한 연구 주제로 대두하였다. 그러나 영어를 대상으로 하는 혐오 표현 연구나 벤치마크 데이터와 대비하여, 한국어를 대상으로 한 혐오 표현의 연구 및 공개 데이터는 부족한 실정이다. 또한 한국어 대상 혐오 표현 연구의 다수가 트위터, 페이스북 등의 SNS 데이터나 인터넷 게시판 데이터를 활용한 것으로, 챗봇과 사용자의 상호작용 관점에서 실사용 데이터를 활용해 혐오 표현 탐지 연구를 수행한 사례는 찾아보기 쉽지 않다.

이 연구는 크라우드소싱으로 수집되어 혐오 표현이 포함된 챗봇-사용자 대화 데이터 35만 개에 한국어로 학습된 트랜스포머 기반 혐오 표현 탐지 모델[2](KoELECTRA Hate Speech Classification Model)을 적용하여, 성능과 한계점을 검토함으로써 개선방안을 논한다. 또한, 기존의 SNS와 댓글에 포함된 혐오 표현 탐지[3,4,5]와는 다른, 챗봇과 사람의 대화문에 포함된 혐오 표현 탐지를 위한 딥러닝 모델링 기법의 방향성을 제시하여, 미래 AI와 사람의 건전한 상호작용 문화에 기여하고자 한다.

2. 관련 연구

온라인상에서 특정 개인이나 그룹을 대상으로 한 혐오 표현 사용 문제가 대두되고 있으며, 이런 혐오 및 욕설 표현은 인신공격, 명예 훼손 등의 형태로 온라인 사용자들에게 피해를 주고 있다. 해당 문제를 해결하려는 방법의 하나로 혐오 및 욕설 표현을 자동 탐지하기 위한 노력이 다양하게 시도되어왔다.

혐오 표현에 관련된 단어 사전을 이용해 분류 모델을 개발하는 연구들이 있었으나, 사전 기반의 모델은 어휘의 변형에 대해 취약하거나 문장의 맥락을 고려하지 못하는 한계점이 존재했다[3,4]. 이러한 사전 기반 모델의 한계를

[1] https://www.dongascience.com/news.php?idx=43063

극복하기 위해 딥러닝 기법을 활용한 연구들도 등장했다. [5]의 연구는 Highway Network 기반 CNN 분류 모델링과 OOV(Out of Vocabulary) 사전학습 임베딩을 통해 인터넷 뉴스 악성 댓글의 편견 및 혐오 표현에 대해 각각 3가지 클래스로 분류했으며 Weighted F1-score 기준 67.49%를 달성했다. 또 다른 연구도[6] 어텐션 기반 다중 채널 CNN 모델링 기법을 적용해 인터넷 뉴스의 악성 댓글을 7가지 혐오 표현 항목으로 이진 분류했고, 가중 평균 F1-score 기준 70.32%의 성능을 달성한 바 있다.

또한 전이학습의 도입을 통해 딥러닝 모델이 더 높은 성능을 달성하도록 할 수 있는 가능성이 드러났다. 2018년 10월에는 트랜스포머 기반의 전이학습을 위한 대규모 언어 모델인 BERT(Bidirectional Encoder Represen-tations from Transformers)[7]가 도입되었다. BERT 는 LSTM을 포함한 기존 모델보다 종합적인 자연 언어처리 문제에서 더 높은 성능을 보여주었다. 2020년 3월에는 비효율적인 BERT의 학습 방식을 보완한 생성 모델 - 판별 모델 구조 기반의 학습법인 ELECTRA(Efficiently Learning an Encoder that Classifies Token Replacements Accurately)[8]가 발표되었다. ELECTRA 모델은 대체 토큰 탐지 방식을 사용하여 모든 토큰에 대해 학습이 가능하여, 모델의 사이즈를 줄이는 동시에 더 높은 성능을 보였다.

KoELECTRA[9]는 ELECTRA의 학습 방식을 한국어 대규모 말뭉치에 적용해 공개된 모델이다. 해당 모델은 기존의 다국어 언어 모델이나 BERT 기반 한국어 언어 모델보다 다양한 문제에서 높은 성능을 보였고, 특히 혐오 표현 판별에서는 F1-score 67.61의 성능을 보고하였다[9]. 또한, BERT-base 기반 모델보다 경량의 모델이기 때문에, Fine-tuning을 통한 비교실험을 수행하기 쉽다. 따라서 혐오 표현 판별 성능과 비교 실험의 수행 용이성을 고려하여, KoELECTRA 모델에 한국어 악성 댓글 데이터를 학습시킨 혐오 표현 탐지 모델[2]을 연구에 활용하였다.

3. 연구 방법론

3.1 데이터

연구를 위한 데이터셋으로 [10,11]에서 공개한 한국어 뉴스 악성 댓글 데이터셋과, 심심이 주식회사가 크라우드소싱을 통해 권리를 취득하여 태깅을 거친 챗봇 학습데이터인 심심이 나쁜말 데이터셋을 활용하였다.

3.1.1 한국어 뉴스 악성 댓글 데이터셋

한국어 뉴스 악성 댓글 데이터셋은 포털 연예 뉴스의 악성 댓글을 수집한 데이터이며[10], 현재까지 공개된 유일한 한국어 차별 및 혐오 발언 데이터셋이다. 데이터는 학습, 검증, 그리고 테스트 데이터로 각각 7,896개, 471개, 974개로 구성되어 있다. 각 데이터 샘플은 댓글 본문과 해당 댓글의 혐오 표현 및 편향 유무를 기술한 라벨로 구성된다. 혐오 표현의 경우 혐오(Hate), 불쾌(Offensive), 일반(None)의 세 가지 클래스로 태깅 되었으며, 특정 대상에 대한 모욕이나 비방이 표현이 포함되면 혐오 라벨로, 냉소나 비꼼 등으로 상대방의 기분을 나쁘게 할 경우 불쾌로, 둘 다 아닐 경우 일반으로 분류되었다. 반면, 편향 라벨은 그 유무와 종류에 따라 다음의 세 가지 클래스로 분류되었다: 성별(Gender, 젠더 관련 편향 포함), 기타(Others, 젠더 외의 정치, 외모, 장애 등에 대한 편향 포함), 일반(None, 편향 미포함).

이 논문에서는 라벨링 된 학습데이터와 검증데이터를 연구에 활용하였고, 라벨 접근이 불가능한 테스트 데이터 대신 검증데이터를 모델 성능 평가에 사용하였다.

3.1.2 심심이 나쁜말 데이터셋

심심이는 2002년 상업용 챗봇 서비스를 시작, 81개 언어로 확장하여 서비스 중인 일상 대화 챗봇 서비스이다. 해당 기업 심심이(주)는 크라우드소싱 방식을 통해 권리를 취득한 1억 건 이상의 챗봇-사람의 대화 데이터를 보유하고 있다. 챗봇(심심이)이 사용자가 원하는 말을 구사할 수 있도록 '내가 이렇게 말하면 심심이는 이렇게 대답한다'는 문답 형태의 대화 짝을 사용자가 직접 챗봇에 가르치게 하고, 해당 '가르치기 데이터'를 챗봇의 대화 DB로 활용 중이다.

심심이 나쁜말 데이터셋은 '가르치기 데이터' 중, 챗봇의 응답 문장만을 대상으로 해당 문장이 서비스의 컨텐츠 규정에 어긋나는 문장(이하 '나쁜말')인지를 크라우드소싱 방식을 통해 모집한 패널에 의해 라벨링 한 데이터다. 각각의 응답 문장은 신뢰도가 검증된 최소 10명의 패널에 의해 평가되었으며, 패널 중 나쁜말 판별 기준에 해당한다고 평가한 패널의 비율이 '나쁜말 점수'로 등록되었다. 예시로 챗봇의 특정 응답에 대해, 10명의 패널 중 4명이 나쁜말에 해당한다고 평가했다면 해당 응답의 나쁜말 점수는 0.4가 된다. 라벨링을 위해 패널에게 제공된 나쁜말의 기준은 다음과 같다. 1) 노골적인 성적 콘텐츠나 음란물 2) 지나친 폭력 또는 기타 위험한 행위를 포함하거나 조장 3) 다른 사람을 괴롭히거나 따돌리는 행위 4) 아동을 성적으로 묘사 5) 인종, 민족성, 피부색, 출신 국가, 종교, 장애 등을 이유로 특정 집단에 대한 증오심을 조장 6) 자연재해, 잔혹 행위, 물리적 충돌, 죽음 또는 기타 비극적인 사건 7) 사기, 도박, 마약, 불법적인 판매.

이 연구에서는 심심이가 보유하고 있는 나쁜말 데이터셋 중 약 5%를 무작위로 뽑아낸 351,432개 데이터를 활용하였으며, 연구 대상 데이터 중 90%를 학습데이터로, 10%를 테스트 데이터로 활용하였다. 또한, 분류기 학습을 진행하기 위해 나쁜말 점수의 1/3, 2/3를 경계로 하여 문장에 각각 상(High), 중(Medium), 하(Low) 라벨을 임의로 부여하였다. 연구에 사용된 데이터의 나쁜말 점수 분포와 예시는 그림 1과 표 1과 같다.

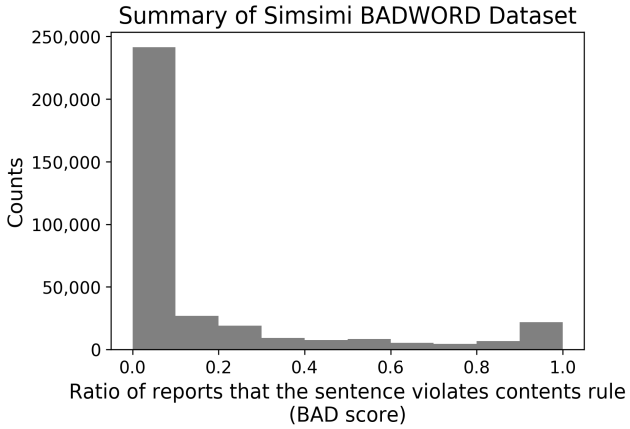

그림 1. 심심이 나쁜말 데이터셋의 나쁜말 점수 분포
(1.0= 명백한 혐오 표현, 0.0 = 비혐오 표현)

표 1. 심심이 나쁜말 데이터셋의 예시.

분류	내용	나쁜말 점수
Low	좋아하는 것이나 취미를 공유해보는 것도 좋습니다	0.0
	아수꾸림	0.2
Medium	죽는게 세상에 도움이 됨	0.5
	아~~그짱구닮은 호구?	0.6
High	아~ 그*신??ㅋㅋㅋㅋ	1.0
	오빠 믿어 오빠가 키스랑 ** 해줄께..	1.0

* 혐오 표현 혹은 일부 성적 표현은 별표로 표시

3.2 분석 방법

이 연구에서는 KoELECTRA[9] 기반의 혐오표현 탐지 모델 KoELECTRA Hate Speech Classification Model[2]을 분석에 사용하였다. 해당 모델은 문장에서의 차별(Bias)와 혐오(Hate)표현 유무를 동시에 예측하는 분류 모델이며, KoELECTRA에 3.1.1.에 언급된 한국어 뉴스 악성 댓글 데이터셋을 학습시켜 만들었다. 특히 이 연구에서는 챗봇의 혐오 표현을 대상으로 하므로, 공개된 모델과 파라미터 중에서도 혐오(Hate) 표현 판별에 대해 Fine-tuning된 koelectra-base-v3-hate-speech 버전의 모델을 활용하였다.

먼저 성능 검증을 위해, 해당 모델에 한국어 뉴스 악성 댓글 데이터셋의 혐오 표현 유무 판별을 Fine-tuning시킨 뒤 그 성능을 확인하였다. 학습에는 크로스 엔트로피가 다중 분류 문제의 손실함수로 활용되었고, 모델이 수렴할 때까지 학습이 진행되었다. 두번째로 심심이 나쁜말 데이터셋을 같은 모델에 동일한 방식과 조건으로 적용해 학습시켰으며, 그 뒤 정량적 평가를 진행하였다. 마지막으로 모델을 통한 나쁜말 판별에 실패한 데이터에 대해 정성 분석을 진행하여 한계점을 분석하고자 하였다.

4. 연구 결과

4.1 모델 성능 정량 분석 결과

각각의 데이터를 KoELECTRA기반 혐오표현 탐지 모델에 적용한 결과는 다음과 같다.

표 2. KoELECTRA기반 혐오 표현 탐지 모델 성능

Data	Accuracy	F1-score
한국어 뉴스 악성 댓글 데이터셋	0.631	0.627
심심이 나쁜말 데이터셋	**0.861**	**0.661**

KoELECTRA 기반 혐오 표현 분류기의 Fine-tuning을 통해 한국어 뉴스 악성 댓글 검증데이터에 대하여 가중 평균 F1-score 0.63을 달성하였다. 이 값은 [5]에서 하이웨이 네트워크 CNN 기반 다중 분류기가 혐오 표현(Hate) 분류에 대해 달성한 0.62에 근접한 높은 성능이다. 또한, 심심이 나쁜말 데이터셋에 대하여 학습시킨 모델은 테스트 데이터에서 가중 평균 F1-score 0.661을 달성하여, KoELECTRA 혐오 표현 탐지 모델이 챗봇-사람의 대화 데이터 맥락에서의 혐오 표현 탐지에도 높은 성능을 보임을 확인할 수 있었다.

그림 2와 3은 판별 결과를 혼동 행렬(Confusion Matrix)로 나타낸 것이다. 한국어 뉴스 악성 댓글 데이터셋에 대해 학습한 모델의 경우 공격적 표현이 포함된 문장을 일반(None)으로 잘못 판별한 경우가 많았다. 그와 반대로 심심이 나쁜말 데이터셋을 학습한 모델의 경우 나쁜말 점수가 High인 문장을 Low로 예측한 결과보다는 나쁜말 점수가 Low인 문장을 High로 더 빈번하게 예측하는 경향을 보였다.

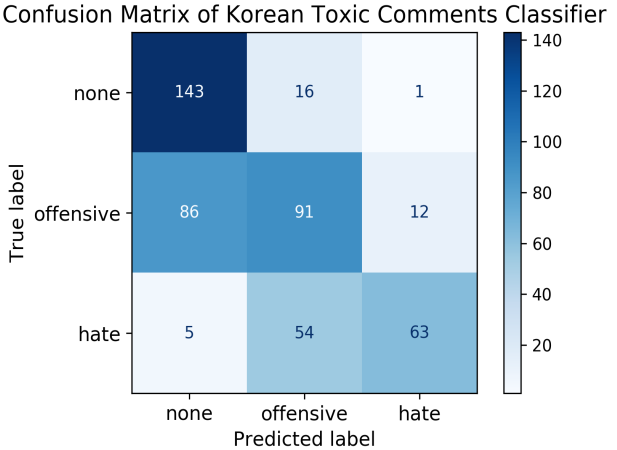

그림 2. 뉴스 악성 댓글 데이터셋에 대해 학습한 모델의 Confusion Matrix

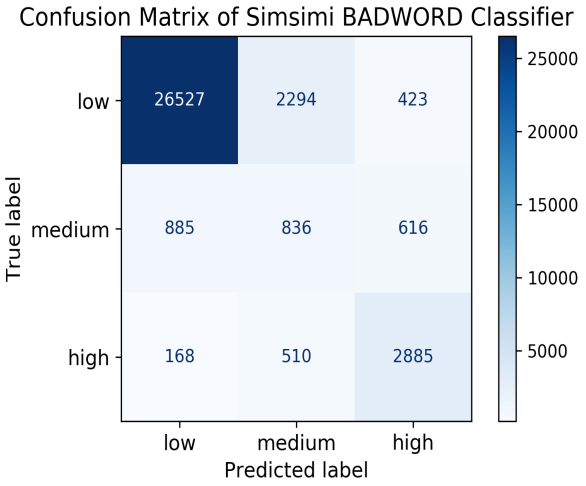

그림 3. 심심이 나쁜말 데이터셋에 대해 학습한 모델의 Confusion Matrix

4.2 정성 분석 결과

분류가 잘못된 데이터(False Positive and False Negative)를 중심으로 정성 분석을 진행해, 한계점을 검토했다.

4.2.1 비정형 데이터

오탈자가 있는 데이터는 그렇지 않은 데이터보다 혐오 표현의 판별이 어려운 것으로 알려져 있다[12]. 특히 Dictionary의 크기가 한정된 모델의 경우 비교적 드물게 나타나는 단어를 '[UNK]' 토큰으로 변환하여 모델에 입력하기 때문에, 특히 비슷하게 소리 나는 문자가 많은 한글 데이터의 경우 판별이 더 어려울 수 있다.

따라서 KoELECTRA 혐오 표현 탐지 모델에 의해 혐오 표현으로 잘 분류된 문장과 잘못 분류된 문장 간의 맞춤법 오류의 개수와 '[UNK]' 토큰의 비율을 확인하였다. 맞춤법 오류의 개수는 Python hanspell 라이브러리를 통해 네이버 맞춤법 검사기[13]를 활용하여 측정하였고, 그 결과는 표 3과 같다.

표 3. 혐오 발언 탐지 모델의 결과 별 평균 오탈자 수와 [UNK] 토큰 비율.

	Average Number of Typo	Average Proportion of [UNK] Token
Accurately classified	1.17	0.015
Inaccurately classified	1.33	0.027

표 3의 결과와 같이, 모델에서 혐오문장으로 탐지하지 못한 문장의 경우 맞춤법 오류가 문장 별 평균 1.33개로, 잘 탐지된 문장들에 포함된 1.17개보다 더 많았다. 또한, 잘못 탐지된 문장이 포함하고 있는 '[UNK]' 토큰의 비율도 2.7%로 잘 탐지된 문장의 '[UNK]' 토큰 비율인 1.5%보다 1.2%p 정도 높았다. 즉 표 4의 예시 사례들과 같이, 사용자가 채팅 맥락에서 띄어쓰기 미사용, 오탈자 사용 및 기호 등으로 단어를 변형해서 사용 시, 토큰화 방식에 따라서 모델을 통한 자동 탐지에서 문장이 잘못 분류될 가능성이 커질 수 있다.

4.2.2. 대화 맥락 미반영 및 소수자 발언에 대한 편향

포털 뉴스 댓글과 달리, 챗봇-사람과의 대화 데이터는 발화문장과 응답 문장을 함께 고려해야 화자들의 의도를 정확히 파악할 수 있는 특징이 있다. 그러나 KoELECTRA 기반의 혐오 표현 분류기와 실험에 사용된 챗봇의 나쁜말 데이터셋은 사용자의 발화 문장을 고려하지 않고, 응답 문장만을 대상으로 했기 때문에, 문맥상으로는 혐오 표현이 아님에도 불구하고 나쁜말 정도가 더 높게 예측되는 사례가 존재했다(표 5). 특히 대화의 맥락이 반영되지 않았기에, [14,15]의 연구 사례에서 보였듯이, 모델이 소수자의 발언을 실제 의미 이상으로 혐오 표현으로 판별하는 편향을 나타내었다.

표 4. 오탈자를 포함한, 잘못 판별된 문장 예시.

Text	나쁜말 점수	Prediction
혀잘라분다	0.5	Low
걍뒤졋이놈아	0.8	Low
하..구랭나랑...사..기..자!♥	0.1	High
ssss eeeeee xxxx	0.4	Low
날 햙하줘	0.8	Low
섂*핛곲싧닳싫*	0.8	Low
ㅗㅗㅗ먹어정*호로씾*갸	1.0	Low

* 혐오 표현 혹은 일부 성적 표현은 별표로 표시

표 5. 맥락 미반영으로 잘못 판별된 문장 및. 소수자 발언에 대한 편향 예시.

응답문장	발화문장	나쁜말 점수	Prediction
게이멋져	게이	0.1	Medium
난 게이에요...	You gay?	0.1	Medium
응 안녕 넌몇살이고... 애인있어? 난 참고로 트랜스젠더야	안녕	0.2	Medium

4.2.3. 부정확한 라벨링

데이터의 라벨이 부정확한 경우 문장이 잘못 분류될 가능성이 컸으며, 학습데이터와 테스트 데이터 모두에서 부정확하게 라벨링 된 데이터들을 발견했다. 특히 심심이 나쁜말 데이터 세트에서는 사회적 편견을 암시하거나, 은어, 신조어가 포함된 문장의 경우 문장의 혐오 정도에 비해 나쁜말 점수가 낮게 평가되는 경향이 발견되었으며 해당 문장들에 대한 모델 판별 결과 역시 비교적 정확하지 못했다.

학습데이터에 포함된 잘못된 라벨은 모델의 성능을 떨어뜨리고, 검증데이터와 테스트 데이터에 포함된 잘못된 라벨은 모델의 정확한 평가를 어렵게 한다. [16]는 전문가가 라벨링 한 혐오 표현 데이터보다 비전문가의 라벨링 한 혐오 표현 데이터는 더 엄격한 기준을 적용해야 모델 성능 저하를 막을 수 있다고 언급한 바 있다.

표 6. 데이터의 라벨이 잘못된 샘플의 예시와, 해당 문장에 대한 판별 결과

Text	나쁜말 점수	Prediction
착짱죽짱	0.1	Medium
한남 빨리 재기해~~	0.0	Low

5. 결론

이 연구는 뉴스 댓글로 학습된 KoELECTRA 기반 혐오 표현 탐지 모델을 챗봇-사람의 대화 데이터에 적용해 이러한 대화 상 혐오 표현에 대한 모델의 분류 성능을 확인하였다. 또한 판별 결과에 대한 정성 분석을 통해 챗봇-사람 대화 맥락에서 혐오 표현 탐지 성능을 저하하는 요인들을 탐색하였다.

그 결과 챗봇-사람의 대화 데이터셋에 포함된 혐오 표현 탐지에서 KoELECTRA 기반 모델이 일반 뉴스 댓글 상의 탐지만큼이나 높은 성능을 보임을 확인하였다. 실험 후 분류가 잘못된 데이터(False Positive and False Negative)를 중심으로 정성 분석을 진행해 한계점을 검토한 결과, 성능 저하의 주요 원인은 다음과 같다.

첫째, 챗봇-사람의 대화 데이터셋은 1:1 대화체 문장들로 댓글이나 SNS 데이터와 비교해 비정형 텍스트를 많이 포함하며, 빈번한 맞춤법 오류와 변형된 단어 사용이 모델의 성능 저하로 이어졌다. 둘째, 대화형 에이전트라는 챗봇의 특성과 혐오 표현의 특성상 대화의 맥락이 문장의 진위 판별에 중요한 반면, 기존의 KoELECTRA 혐오 표현 분류 모델에서는 발화 문장을 제외한 응답 문장만을 학습시키므로, 대화 세트의 맥락을 학습시키지 못했다는 한계가 있었다. 이로 인해 소수자 발언을 실제 의미 이상으로 혐오 표현으로 판별하는 편향을 나타냈다. 셋째, 데이터 라벨링의 부정확성 이슈가 있었다. 욕설의 강도에 대한 평가는 개인의 욕설에 노출되는 정도와 성향에 따라 다르다[17]. 따라서 특정 혐오 표현에 대해 누군가는 심한 욕설이라고 평가하고, 누군가는 욕설이 아니라고 평가할 가능성이 존재한다. 크라우드소싱 방식으로 라벨을 진행한 뉴스 댓글 데이터 및 심심이 나쁜말 데이터셋에서 부정확하게 라벨링 된 데이터들이 발견되었으며, 이는 모델의 성능을 저하시켰다.

그러나 본 논문은 온라인 뉴스 악성 댓글만 학습시킨 KoELECTRA Hate Speech 분류 모델의 챗봇에 포함된 혐오 표현 분류 성능과 한계점 분석에만 초점을 맞추었다는 한계도 존재한다. 향후 연구에서는 KoELECTRA 혐오 표현 분류기뿐 아니라 타 혐오 표현 탐지 모델과의 비교연구를 통해 모델 간 차이와 특성에 대한 검증을 거칠 수 있을 것이다.

이번 연구를 통해 딥러닝을 활용한 챗봇-사람 대화 데이터셋에 포함된 혐오 표현 탐지의 한계점에 대한 개선 방향을 다음과 같이 제시한다. 먼저 혐오 표현 탐지에서 채팅 맥락 상 빈번히 사용되는 비정형 텍스트의 패턴을 적극 활용하는 노력이 필요하다. 그 예로 비정형 텍스트로 변형된 문장을 더욱 명확한 의미의 문장으로 바꾸는 채팅 맥락을 고려한 전처리 작업을 고려할 수 있다. 다음으로, 향후 연구에서는 발화문장-응답 문장을 함께 학습시켜 대화의 맥락을 반영하고, 이를 통해 혐오 표현 탐지 모델의 성능을 향상할 수 있다. 마지막으로, 다양한 이용자군으로부터 수집된 혐오 표현 라벨 데이터셋을 확보하는 방법이 있다. 이와 더불어 단순 혐오 표현의 유무만이 아닌 차별과 혐오의 대상, 혐오 표현의 분류를 기술한 세부적인 라벨링 확보가 학습에 도움을 주리라 기대한다. 심심이 나쁜말 데이터셋과 한국어 뉴스 악성 댓글의 데이터들은 크라우드 방식으로 모인 패널들이 제시된 기준에 따라 나쁜말 유무와 강도의 정도를 평가하는 방식으로 라벨링 되었다. 현재 평가 방식에 전문가 검수 단계를 추가하거나, 딥러닝 모델을 통해 자동 분류된 결과에 대해서 패널이 추가 평가하는 방식 등을 적용해 라벨링의 정확도를 높일 수 있다. 이러한 다양한 노력은 혐오 표현 탐지 모델의 성능 향상에 기여할 수 있을 것이다.

사사문구

이 논문은 과학기술정보통신부의 재원으로 기초과학연구원의 지원(IBS-R029-C2)과 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2017R1E1A1A01076400).

참고문헌

[1] K. Darling, “Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects,” In *Robot law,* Edward Elgar Publishing, 2016.

[2] J. Park, “Korean Hate Speech Classification,” Github repository, https://github.com/monologg/korean-hate-speech-koelectra
[3] 김현정, 윤영미, 이병문, “향상된 FFP (Feature Frequency Profile) 을 활용한 악성 댓글의 판별시스템”, 한국정보기술학회논문지, 9(1), pp.207-216, 2011.
[4] 홍진주, 김세한, 박제원, 최재현, “감성분석 SVM 을 이용한 인터넷 악성댓글 탐지 기법”, 한국정보통신학회논문지, 20(2), pp.260-267, 2016.
[5] 이현상, 이희준, 오세환, “딥러닝 기술을 활용한 악성댓글 분류: Highway Network 기반 CNN 모델링 연구”, 한국경영학회 통합학술발표논문집, pp.343-351, 2020.
[6] 이원석, & 이현상. (2020). 딥러닝 기술을 활용한 차별 및 혐오 표현 탐지: 어텐션 기반 다중 채널 CNN 모델링. 한국정보통신학회논문지, 24(12), 1595-1603.
[7] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding”, *arXiv preprint, arXiv:1810.04805*, 2018.
[8] K. Clark, M. T. Luong, Q. V. Le and C. D. Manning, “ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators,” *ICLR*, 2020.
[9] J. Park, “KoELECTRA: Pretrained ELECTRA Model for Korean,” Github repository, https://github.com/monologg/KoELECTRA, 2020.
[10] J. H. Moon, W. I. Cho, and J. B. Lee, “Beep! Korean Corpus of Online News Comments for Toxic Speech Detection,” *Proceeding of the 8th International Workshop on Natural Language Processing for Social Media*, Taipei, 2020.
[11] 조원익, 문지형, “한국어 혐오 표현 코퍼스 구축 방법론 연구: 온라인 악성 댓글에 나타나는 특성을 중심으로”, 제32회 한글 및 한국어 정보처리 학술대회 논문집, pp.298-303, 2020.
[12] T. Gröndahl, L. Pajola, M. Juuti, M. Conti, and N. Asokan, “All You Need is "Love": Evading Hate Speech Detection,” *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security* (*AISec '18*), pp. 2-12, 2018.
[13] "py-hanspell,” Github repository, https://github.com/ssut/py-hanspell
[14] T. Davidson, D. Bhattacharya, and I. Weber, “Racial Bias in Hate Speech and Abusive Language Detection Datasets,” *Proceedings of the Third Workshop on Abusive Language Online*, pp. 25-35, 2019.
[15] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, “The risk of racial bias in hate speech detection,”, *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1668-1678, 2019.
[16] Z. Waseem, “Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter,” *Proceedings of the first workshop on NLP and computational social science*, pp. 138-142, 2016.
[17] D. A. Infante, B. L. Riddle, C. L. Horvath, and S.-A. Tumlin, “Verbal aggressiveness: Messages and reasons,” *Communication Quarterly* 40, 2, pp. 116–126. 1992.